

Constructions for Clumps Statistics

F. Bassino¹, J. Clément², J. Fayolle³, and P. Nicodème⁴

¹IGM, Université de Marne la Vallée, 77454 Marne-la-Vallée Cedex 2, France. Frederique.Bassino@univ-mlv.fr

²GREYC, CNRS-UMR 6072, Université de Caen, 14032 Caen, France. Julien.Clement@info.unicaen.fr

³LRI; Univ. Paris-Sud, CNRS ; Bât 490, 91405 Orsay, France. Julien.Fayolle@lri.fr

⁴LIX, CNRS-UMR 7161, École polytechnique, 91128, Palaiseau, France. nicodeme@lix.polytechnique.fr

We consider a component of the word statistics known as clump; starting from a finite set of words, clumps are maximal overlapping sets of these occurrences. This parameter has first been studied by Schbath [22] with the aim of counting the number of occurrences of words in random texts. Later work with similar probabilistic approach used the Chen-Stein approximation for a compound Poisson distribution, where the number of clumps follows a law close to Poisson. Presently there is no combinatorial counterpart to this approach, and we fill the gap here. We emphasize the fact that, in contrast with the probabilistic approach which only provides asymptotic results, the combinatorial approach provides exact results that are useful when considering short sequences.

1 Introduction

Counting words and motifs in random texts has provided extended studies with theoretical and practical reasons. Much of the present combinatorial research has built over the work of Guibas and Odlyzko [10, 11] who defined the autocorrelation polynomial of a word. As an apparently surprising consequence of their work, the waiting time for the first occurrence of the word 111 in a Bernoulli string with probability $1/2$ for zeroes and ones is larger than the waiting time for the first occurrence of the word 100. This is due to the fact that the words 111 occur by *clumps* of ones, the probability of extending a clump by one position being $1/2$; this implies that the average number of 111 in a clump is larger than one; in contrast, there is only one 100 in each clump of 100. Since the probability that the word 111 and the word 100 start at a given position both are $1/8$, the interarrival time of clumps of 111 is larger than the interarrival time of clumps of 100.

We analyze in this article several statistics connected to clumps of one word or of a reduced set of words. Our approach is based on properties of the Régnier-Szpankowski [18] decomposition of languages along occurrences of the considered word or set of words and on properties of the prefix codes generating the clumps. We provide explicit generating functions in the Bernoulli model for statistics such as (i) the number of clumps, (ii) the number of k -clumps, (iii) the number of positions of the texts covered by clumps, and (iv) the size of clumps in infinite texts; these results may be extended to a Markov model, providing some technicalities. We consider also in the Bernoulli model an algorithmic approach where we construct deterministic finite automatas recognizing clumps. This approach extends directly to the Markov model, and we obtain as a direct consequence a Gaussian limit law for the number of clumps in random texts.

Consider a rough first approximation for clumps of one word. If the probability occurrence of a word w is small, the probability of clumps \mathcal{R} of this word is small. This implies that the number of clumps in texts of size n follows a Poisson law of parameter $\lambda = n \times \mathbf{P}(\text{a clump starts at position } i)$, where i is a random position. Approximating further, the random number of occurrences Ω of the word w in a clump follows a geometric law with parameter ω , where ω is the probability of self-overlap of the word. Schbath and Reinert [19] obtained in the Markov case of any order a compound Poisson limit law for the count of number of occurrences by the Chein-Stein method. See Reinert *et al.* [20] for a review and

subm. to DMTCS © by the authors Discrete Mathematics and Theoretical Computer Science (DMTCS), Nancy, France

Barbour *et al.* [1] for an extensive introduction to the Poisson approximation. Schbath [22] give the first moment of the number of k -clumps and of the number of clumps in Bernoulli texts. Recently, Stefanov *et al.* [24] use a stopping time method to compute the distribution of clumps; their results are not explicit and practical application of their method requires the inversion of a probability generating function.

We describe in Section 2 our notations and the Régnier-Szpankowski language decomposition. Section 3.2 and Section 3.3 respectively provide our analysis in the case of counting clumps and k -clumps of one word and of a finite set of words. We prove by an automaton construction a normal limit law for the number of clumps in Section 5

2 Preliminaries

We consider a finite alphabet \mathcal{A} . Unless explicitly stated when considering a Markov source, the texts are generated by a non-uniform Bernoulli source over the alphabet \mathcal{A} . Given a set of words, clumps of these words may be seen as a generalization of runs of one letter.

Clumps and k -clumps. When considering a reduced set of words $U = \{u_1, \dots, u_r\}$ where each word u_i has size at least 2, a clump is a maximal set of occurrences of words of U such that

- any two consecutive letters of the clump belong to (is a factor of) at least one occurrence,
- either the clump is composed of a single occurrence that overlaps no other occurrences, or each occurrence *overlaps* at least one other occurrence.

This definition naturally applies also to the case where U is composed of a single word.

As example, considering the set $U = \{aba, bba\}$ and the text $T = bbbababababbbbabaababb$, we have

$$T = bbbababababbbbabaababb$$

where the clumps are underlined. The word $bbababababa$ beginning at the second position of the text is a clump, and so are the words $bbaba$ and aba beginning at the 15th and 20th positions. On the contrary, the word $ababa$ beginning at the sixth position is not a clump since it is not maximal; neither is a clump the word $bbabaaba$ beginning at the 15th position, since its two-letters factor aa is neither a factor of an occurrence of aba nor of an occurrence of bba .

More formally, we use as an intermediate step *clusters*, following Goulden and Jackson [9].

Definition 1 (Clumps) A clustering-word for the set $\mathcal{U} = \{u_1, \dots, u_r\}$ is a word $w \in \mathcal{A}^*$ such that any two consecutive positions in w are covered by the same occurrence in w of a word $u \in \mathcal{U}$. The position i of the word w is covered by a word u if $u = w[(j - |u| + 1) \dots j]$ for some $j \in \{|u|, \dots, n\}$ and $j - |u| + 1 \leq i \leq j$. A cluster of a clustering-word w in $\mathcal{K}_{\mathcal{U}}$ is a set of occurrence positions subsets $\{S_u \subset \text{Occ}(u, w) \mid u \in \mathcal{U}\}$ which covers exactly w , that is, every two consecutive positions i and $i + 1$ in w are covered by at least one same occurrence of some $u \in \mathcal{U}$. More formally

$$\forall i \in \{1, \dots, |w| - 1\} \quad \exists u \in \mathcal{U}, \exists p \in S_u \quad \text{such that} \quad p - |u| + 1 < i + 1 \leq p.$$

A clump, generically denoted here by \mathfrak{K} is a maximal cluster in the sense that there exists no occurrence of the set \mathcal{U} that overlaps the corresponding clustering word without being a factor of it.

Note that a single word is a cluster and that, as mentionned previously, a clump may be composed of a single word.

A k -clump of occurrences of U (denoted by $\mathfrak{K}^{(k)}$) is a clump containing exactly k occurrences of U .

We aim here at providing explicit analytic formulas for the moments of the number of clumps, the total size of text covered by clumps or the number of clumps with exactly k occurrences.

Notations. We consider the *residual* language $\mathcal{D} = \mathcal{L}.w^-$ as $\mathcal{D} = \{x, x.w \in \mathcal{L}\}$.

In case of ambiguity, we will use a bracket notation $\{\mathcal{L}\}(z, \dots)$ to represent the generating function of the language $\{\mathcal{L}\}$; in particular, for $\mathcal{D} = \mathcal{L}.w^-$, we write $\{\mathcal{L}.w^-\}(z, \dots) = \mathcal{D}(z, \dots)$.

Considering two languages \mathcal{L}_1 and \mathcal{L}_2 , if we have $\mathcal{L}_1 \subset \mathcal{L}_2$, we write $\mathcal{L}_2 - \mathcal{L}_1 = \mathcal{L}_2 \setminus \mathcal{L}_1$ as the difference of sets;

Reduced set of words. A set of words $U = \{u_1, \dots, u_r\}$ is reduced if no u_i is factor of a u_j with i different of j .

Autocorrelations, correlations and right extension sets of words. We recall here the definition of *Right Extension Set* introduced in Bassino *et al.* [2].

The right extension set of a pair of words (h_1, h_2) is

$$\mathcal{E}_{h_1, h_2} = \{e \mid \text{there exists } e' \in \mathcal{A}^+ \text{ such that } h_1 e = e' h_2 \text{ with } 0 < |e| < |h_2|\}.$$

If the word h_1 is not factor of h_2 this extension set of h_1 to h_2 is the usual correlation set of h_1 and h_2

When we have $h_1 = h_2$, we get the autocorrelation set $\mathcal{C}_{h,h}$ of the word h that we will note further \mathcal{C} when there is no ambiguity.

We also note $\mathcal{C}_\circ = \mathcal{C} - \epsilon$. Remark that \mathcal{C}_\circ is empty if the word w has no autocorrelation.

We remark here that the empty word ϵ belongs to the autocorrelation set of a word. Note also that the correlation set of two words may be empty.

We have as examples

$$\mathcal{C}_{aaba, aab} = \{b, ab\}, \quad \mathcal{C}_{ababa, ababa} = \{\epsilon, ba, baba\}.$$

Generating functions. We aim at computing the number of a given object in random texts by use of generating functions such as

$$L_v(z, x) = \sum_{T \in \mathcal{L}} \mathbf{P}(T) z^{|T|} x^{|T|_v} = \sum l_{n,i} x^i z^n \quad (1)$$

where $|T|_v$ is the number of occurrences of the object v in the text T and $l_{n,i}$ is the probability that a text of size n has i occurrences of this object. This extends naturally for counting more than one object by considering multivariate generating functions with several parameters.

If the random variable X_n counts the number of objects in a text of size n , we get from Equation (1)

$$\mathbf{E}(X_n) = [z^n] \frac{\partial L(z, x)}{\partial x} \Big|_{x=1}, \quad \mathbf{E}(X_n^2) = [z^n] \frac{\partial}{\partial x} x \frac{\partial L(z, x)}{\partial x} \Big|_{x=1}.$$

Recovering exactly or asymptotically these moments follows then from classical methods.

3 Formal language approach

3.1 Régnier and Szpankowski decomposition

Since our work extends the formal language approach of Régnier and Szpankowski [18], we recall it here.

Considering one word w , Régnier and Szpankowski use a natural parsing or decomposition of texts with at least one occurrence of w , where

- there is a first occurrence at the right extremity of a “subtext”, the set of which constitute a *Right* language,
- possibly followed by other occurrences, that are separated by “subtexts” that constitute the *Minimal* language,

- and completed by “subtexts” that provide no other occurrences.

Moreover, there is a language without any match of the considered word w . Régnier [17], further extended this approach to a reduced set of words.

We follow here the book of Lothaire [15](Chapter 7) which presents their method.

We consider a set of words $V = \{v_1, \dots, v_r\}$. We have, formally

Definition 2 *Right, Minimal, Ultimate and Not languages*

- The “Right” language \mathcal{R}_i associated to the word v_i is the set of words

$$\mathcal{R}_i = \{r \mid r = e.v_i \text{ and } \nexists e' \in V, r = xe'y, |y| > 0\}.$$

- The “Minimal” language \mathcal{M}_{ij} leading from a word v_i to a word v_j is the set of words

$$\mathcal{M}_{ij} = \{m \mid v_i.m = e.v_j \text{ and } \nexists e' \in V, v_i.m = xe'y, |x| > 0, |y| > 0\}.$$

- The “Ultimate” language completing a text after an occurrence of the word v_i is the set of words

$$\mathcal{U}_i = \{u \mid \nexists e \in V, v_i.u = xey, |x| > 0\}.$$

- The “Not” language completing a text after an occurrence of the word v_i is the set of words

$$\mathcal{N} = \{n \mid \nexists e \in V, n = xey\}.$$

The notations \mathcal{R} , \mathcal{M} , \mathcal{U} and \mathcal{N} refer here to the Right, Minimal, Ultimate and Not languages of a single word.

Considering as example the word $w = ababa$; in the following texts, the underlined words belong to the set \mathcal{M} ; the overlined text does not since the word represented in bold faces is an intermediate occurrence.

$$ababa\underline{aaaaababa} \quad ababa\overline{ababababbbbbb} \quad ababa\underline{ba}.$$

Considering the matrix \mathbb{M} such that $\mathbb{M}_{ij} = \mathcal{M}_{ij}$, we have

$$\bigcup_{k \geq 1} (\mathbb{M}^k)_{i,j} = \mathcal{A}^* \cdot w_j + \mathcal{C}_{ij} - \delta_{ij}\epsilon, \quad \mathcal{U}_i \cdot \mathcal{A} = \bigcup_j \mathcal{M}_{ij} + \mathcal{U}_i - \epsilon, \quad (2)$$

$$\mathcal{A} \cdot \mathcal{R}_j - (\mathcal{R}_j - w_j) = \bigcup_i w_i \mathcal{M}_{ij}, \quad \mathcal{N} \cdot w_j = \mathcal{R}_j + \bigcup_i \mathcal{R}_i (\mathcal{C}_{ij} - \delta_{ij}\epsilon). \quad (3)$$

If the size of the texts is counted by the variable z and the occurrences of the words v_1, \dots, v_r are counted respectively by x_1, \dots, x_r , we get the matrix equation

$$F(z, x_1, \dots, x_r) = \mathcal{N}(z) + (x_1 \mathcal{R}_1(z), \dots, x_r \mathcal{R}_r(z)) (\mathbf{I} - \mathbb{M}(z, x_1, \dots, x_r))^{-1} \begin{pmatrix} \mathcal{U}_1(z) \\ \vdots \\ \mathcal{U}_r(z) \end{pmatrix}. \quad (4)$$

In this last equation, we have $\mathbb{M}_{ij}(z, x_1, \dots, x_r) = x_j \mathcal{M}_{ij}(z)$ and the generating functions $\mathcal{R}_i(z)$, $\mathcal{M}_{ij}(z)$, $\mathcal{U}_j(z)$ and $\mathcal{N}(z)$ can be computed explicitly from the set of Equations (2, 3).

In particular, when considering the Bernoulli weighted case $A(z) = z$ and a single word w with $\pi_w = \mathbf{P}(w)$, we have the set of equations

$$R(z) = \frac{\pi_w z^{|w|}}{D(z)}, \quad M(z) = 1 + \frac{z-1}{D(z)}, \quad U(z) = \frac{1}{D(z)}, \quad N(z) = \frac{C(z)}{D(z)} \quad \left(\frac{1}{D(z)} = \frac{1}{\pi_w z^{|w|} + (1-z)C(z)} \right) \quad (5)$$

$$\mathcal{A}^* = \mathcal{N} + \mathcal{R}\mathcal{M}^*\mathcal{U} \implies F(z, x) = \frac{1}{1 - z + \pi_w z^{|w|} \frac{1-x}{x + (1-x)C(z)}} = \sum_{n,k} f_{n,k} x^k z^n. \quad (6)$$

In this last equation, $f_{n,k}$ is the probability that a text of size n has k occurrences of w .

3.2 Clump analysis for one word

The decomposition of Régnier and Szpankowski is based on a parsing by the occurrences of the considered words. We use a similar approach, but parse with respect to the occurrences of clumps. As a major difference, when they consider the minimal language separating two occurrences, these two occurrences may overlap; in contrast, by definition, overlapping of clumps is forbidden.

A clump of the word w is basically defined as $w\mathcal{C}^*$, since any element of \mathcal{C}_o concatenated to a cluster extends this cluster.

Considering the word $w = aaa$, we have $\mathcal{C} = \{\epsilon, a, aa\}$ and \mathcal{C}^* is ambiguous. We can however generate unambiguously \mathcal{C}^* as described in the next section.

3.2.1 A prefix code \mathcal{K} to generate unambiguously \mathcal{C}^*

Since \mathcal{C}_o is a finite language, it is possible to find a prefix code \mathcal{K} generating \mathcal{C}_o ; moreover, for $c_1, c_2 \in \mathcal{C} - \epsilon$ and $|c_1| < |c_2|$, the word c_1 is a proper suffix of c_2 . Otherwise stated, the prefix code $\mathcal{K} = \{\kappa_1, \dots, \kappa_k\}$ is built over words q_1, q_2, \dots, q_k and may be written as $\mathcal{K} = \{q_1, q_2 q_1, \dots, q_k q_{k-1} \dots q_1\}$.

We Refer to Berstel and Perrin [4] for an introduction to prefix codes. See also Berstel [3] for an analysis of counts of words of the pattern U by semaphore codes $U = \mathcal{A}^* U \mathcal{A}^+$. We have the following lemma

Lemma 1 *The prefix code $\mathcal{K} = \mathcal{C}_o \setminus \mathcal{C}_o \mathcal{A}^+$ generates unambiguously the language \mathcal{C}^* .*

Proof: It is clear that \mathcal{K} is prefix. Consider $w \in \mathcal{C}_o - \mathcal{K}$ if this last set is not empty. Since \mathcal{K} is a set of words of \mathcal{C} without any prefix in \mathcal{C} , we have a *contrario* $w = u.v$ with u and v non-empty and in \mathcal{C} . We have $|u| < |w|$ and $|v| < |w|$; if u or v does not belong to \mathcal{K} , we may iterate the process on the corresponding word. Since $|w|$ is finite, after a finite number of steps, we get to a decomposition $w = \kappa_{i_1} \dots \kappa_{i_j}$ where each κ_{i_i} is in \mathcal{K} . Since \mathcal{K} is a code, the decomposition of each word of \mathcal{C} over \mathcal{K} is unique and so is the decomposition of any word of \mathcal{C}^* . \square

Example 1 *Let $w = abaabaaba$. We have*

$$\begin{array}{l} abaabaaba|\epsilon \\ abaaba|aba \\ aba|abaaba \\ a|baabaaba \end{array} \implies \mathcal{C} = \{\epsilon, aba, abaaba, baabaaba\} \implies \mathcal{K} = \{aba, baabaaba\}.$$

The *periods* of a word w is the set of integers $\{|h|, h \in \mathcal{C}_o\}$; the irreducible periods is the subset of periods of which all the periods may be deduced. As follows from Guibas and Odlyzko [10] and Rivals and Rahmann [21], when considering the word $ababaccababa$, the irreducible periods are 7, 9 while the period 11 can be deduced from the periods 7 and 9. However, we have here $\mathcal{K} = \mathcal{C} = \{ccababa, baccababa, babaccababa\}$, which implies somehow against intuition that, in general, there is no bijection between the irreducible periods and the prefix code of a word.

Constructing the prefix-code \mathcal{K} . We use the following algorithm:

1. start with the word w ;
2. shift w to the right to the first self-overlapping position; Let κ_1 be the trailing suffix so obtained; insert it in a trie \mathcal{E} ;
3. repeat shifting, obtaining new trailing suffixes; for each new suffix generated, try an insertion in the trie. If you reach a leaf, drop the suffix; elsewhere insert it.

The worst case complexity for this construction is $O(|w|)$, but the average complexity is $O(|\mathcal{K}| \log(|\mathcal{K}|))$, the average path length of a trie built over $|\mathcal{K}|$ keys.

3.2.2 The language decomposition

Considering the word $w = aaaaa$, we have $\mathcal{C} = \{a, aa, aaa, aaaa\}$ and $\mathcal{K} = \{a\}$. Moreover, we have $\mathcal{M} = \{a, b(b + ab + aab + aaab + aaaaab)^* aaaaa\}$. We get here $\mathcal{K} \subset \mathcal{M}$ and $\mathcal{M} - \mathcal{K} = \mathcal{L}w$; The language \mathcal{M} and \mathcal{K} are indeed connected by a simple property that we describe now.

Lemma 2 *For any word w with autocorrelation set \mathcal{C} , prefix code \mathcal{K} generating \mathcal{C}^* and minimal language \mathcal{M} , there exists a non-empty language \mathcal{L} such that*

$$\mathcal{K} \subset \mathcal{M} \quad \text{and} \quad \mathcal{M} - \mathcal{K} = \mathcal{L}w. \quad (7)$$

Proof: We have $\mathcal{K} \subset \mathcal{C}$ and $\mathcal{K} \subset \mathcal{M}$; therefore, we have $\mathcal{K} \subset \mathcal{M} \cap \mathcal{C}$. We prove that if $w \in \mathcal{C} - \mathcal{K}$ then $w \notin \mathcal{M}$. Let us suppose that $w \neq \epsilon$ and $w \in \mathcal{C} - \mathcal{K}$. This implies that $w \in \mathcal{K}\mathcal{A}^*$ by definition of \mathcal{K} . Therefore, we have $w = \kappa v$ with $\kappa \in \mathcal{K}$ and $|v| > 0$. As a consequence, w cannot belong to the minimal language \mathcal{M} , the word κ corresponding to a previous occurrence of w . \square

This leads immediately to the fundamental lemma.

Lemma 3 *The basic equation for the combinatorial decomposition of texts on the alphabet \mathcal{A} where v counts some object in the clump of a word w is*

$$\mathcal{A}_v^* = \mathcal{N} + \mathcal{R}w^-(w\mathcal{C}^*)_v((\mathcal{M} - \mathcal{K})w^-(w\mathcal{C}^*)_v)^*\mathcal{U}, \quad (8)$$

Proof: The Equation (8) follows from the parsing

- either there is no occurrence of w , the Not language \mathcal{N} ,
- or
 1. we read until the first occurrence : $\mathcal{R}w^-w$,
 2. followed by any number of overlapping occurrences of w (a clump less the first occurrence): \mathcal{C}^* ,
 3. followed by any number of
 - (a) next occurrence of w without overlap: $(\mathcal{M} - \mathcal{K})w^-w$
 - (b) and any number of overlapping occurrences of w : \mathcal{C}^* .

\square

We can now use the preceeding lemma to count several parameters related to the clumps.

3.2.3 Counting parameters related to the clumps

Let $\mathfrak{R}(z, x, t)$ be the generating function where the variable x counts the number of occurrence of w in a clump, and the variable t counts the size of the clumps; the variable z is used here to count the total length of the texts. We also use a variable u to count the number of clumps. We have the following theorem

Theorem 1 *In the weighted model such that $A(z) = z$, the generating function counting the number of occurrences of a word w and the number of positions covered by the clumps of w verifies*

$$F(z, \mathfrak{R}(z, x, t)) = \mathcal{N}(z) + \frac{\mathcal{R}(z)}{\pi_w z^{|w|}} \mathfrak{R}(zt, x) \frac{1}{1 - \frac{\mathcal{M}(z) - \mathcal{K}(z)}{\pi_w z^{|w|}} \times \mathfrak{R}(zt, x)} \mathcal{U}(z) \quad (9)$$

where the generating function of the clumps verifies

$$\mathfrak{R}(z, x, t) = x \pi_w (zt)^{|w|} \frac{1}{1 - x \mathcal{K}(zt)} \quad (10)$$

As a consequence, the generating function counting also the number of clumps is

$$G(z, x, t, u) = F(z, u \mathfrak{R}(z, x, t)). \quad (11)$$

Proof: This theorem follows from Lemma (1) and from a direct translation of Equation (8) into generating functions. \square

3.2.4 Occurrences of clumps.

Considering $G(z, u \mathfrak{R}(z, 1, 1))$ in Equation (9) and using Equation (10) provides the generating function

$$O^{(\gamma)}(z, u) = \sum_{n,i} o_{n,i}^{(\gamma)} u^i z^n = \mathcal{N}(z) + \frac{u \mathcal{R}(z) \mathcal{U}(z)}{1 - u \mathcal{M}(z) + (u-1) \mathcal{K}(z)} \quad (12)$$

where $o_{n,i}^{(\gamma)}$ is the probability of getting i clumps (of any size) in a text of size n . Considering Γ_n , the expectation of number of clumps in texts of size n , we get by differentiation

$$\sum_n \Gamma_n z^n = \frac{\mathcal{R}(z) \mathcal{U}(z) (1 - \mathcal{K}(z))}{(1 - \mathcal{M}(z))^2} = \frac{\pi_w z^{|w|} (1 - \mathcal{K}(z))}{(1 - z)^2}.$$

This implies that $\Gamma_n = (n - |w| + 1) \pi_w (1 - \mathcal{K}(1)) - \pi_w \mathcal{K}'(1)$, to compare with the expectation $(n - |w| + 1) \pi_w$ of the number of occurrences of the word w .

3.2.5 Occurrences of k -clumps.

By considering the equation of a clump of occurrences of w , we can write

$$w \mathcal{C}^* = w + w \mathcal{K} + w \mathcal{K}^2 + \dots (v-1+1) w \mathcal{K}^{k-1} + \dots$$

to count clumps with exactly k occurrences of w .

Writing $\mathfrak{R}^{(k)}(z, v)$ the generating function which counts with the variable z the size of the clumps and where the variable v selects k -clumps, we have

$$\mathfrak{R}^{(k)}(z, v) = \pi_w z^{|w|} \left(\frac{1}{1 - \mathcal{K}(z)} + (v-1) \mathcal{K}(z)^{k-1} \right)$$

Substituting this in Equation 9 gives

$$O^{(\gamma_k)}(z, v) = \sum_{n,i} o_{n,i}^{(\gamma_k)} v^i z^n = \mathcal{N}(z) + \frac{\mathcal{R}(z)}{\pi_w z^{|w|}} \mathfrak{R}^{(k)}(z, v) \frac{1}{1 - \frac{\mathcal{M}(z) - \mathcal{K}(z)}{\pi_w z^{|w|}} \times \mathfrak{R}^{(k)}(z, v)} \mathcal{U}(z),$$

where $o_{n,i}^{(\gamma_k)}$ is the probability that a text of size n contains exactly i k -clumps.

3.2.6 Probability that a random position is covered by a clump

This follows from the knowledge of the number of positions of the texts covered by the clumps.

Let P_n be the random variable counting the number of positions covered by the clumps of a word w in texts of size n and H_n be the probability that a random position is covered by a clump in a text of size n .

Let $F(z, t) = G(z, \mathfrak{R}(zt, 1))$ where $G(z, \mathfrak{R})$ is given by Equation (9) be the generating function counting the size of the texts and the number of positions covered by clumps. We have

$$H_n = \sum_{i \geq 0} \frac{i}{n} \mathbf{P}(P_n = i) \iff H_n = [z^n] \frac{\partial}{\partial t} z \int_0^z F(y, t) dy \Big|_{t=1}.$$

3.3 Clumps of a finite set of words

We provide in this section a matricial solution for counting clumps of a reduced finite set of words. For simplicity sake we consider a set of two words w_1 and w_2 but our approach is amenable to any reduced finite set.

Similarly to the one word case, we are lead to consider prefix codes generating the correlation of two words. Writing \mathcal{C}_{ij}^* with $i \neq j$ makes no sense in terms of language decomposition. However, we can write as previously $\mathcal{K}_{ij} = \mathcal{C}_{ij} - \mathcal{C}_{ij}\mathcal{A}^+$, which defines a minimal correlation language with good properties.

We have as examples

Example 2 Let $w_1 = aabaa$ and $w_2 = aaa$. We have $\mathcal{C}_{12} = \{a, aa\}$ and $\mathcal{K}_{12} = \{a\}$. In this case, we have $\mathcal{C}_{12} = \mathcal{C}_{22} - \{\epsilon\}$ and $\mathcal{K}_{12} = \mathcal{K}_{22}$.

Example 3 Let $w_1 = abab$ and $w_2 = baba$. We have $\mathcal{C}_{12} = \mathcal{K}_{12} = \{a, aba\}$. In this case, we have $\mathcal{C}_{22} = \{\epsilon, ba\}$ and $\mathcal{K}_{12} = a.\mathcal{K}_{22}$.

Following a proof similar to the proof of Lemma (2), there exists a language \mathcal{L} such that

$$\mathcal{M}_{ij} - \mathcal{K}_{ij} = \mathcal{L}.w_j.$$

We can therefore write a minimal correlation matrix \mathbb{K} , consider the matrix $\mathbb{S} = \mathbb{K}^*$ and write a clump matrix \mathbb{G} as follows

$$\mathbb{K} = \begin{pmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{21} & \mathcal{K}_{22} \end{pmatrix}, \quad \mathbb{S} = \mathbb{K}^*, \quad \mathbb{G} = \begin{pmatrix} w_1\mathbb{S}_{11} & w_1\mathbb{S}_{12} \\ w_2\mathbb{S}_{21} & w_2\mathbb{S}_{22} \end{pmatrix} \quad (13)$$

In this equation, \mathbb{G}_{ij} is a clump starting with the word w_i and finishing with the word w_j . We obtain now a fundamental matricial decomposition that can be used for further analysis,

$$\mathcal{A}^* = (\mathcal{R}_1 w_1^-, \mathcal{R}_2 w_2^-) \mathbb{G} \left((\mathbb{M} - \mathbb{K})^- \mathbb{G} \right)^* \begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{pmatrix}$$

where we have $(\mathbb{M} - \mathbb{K})_{ij}^- = (\mathcal{M}_{ij} - \mathcal{K}_{ij})w_j^-$.

4 Automaton approach

For a set $U = \{u_1, \dots, u_r\}$, we build a kind of ‘‘Aho-Corasick’’ automaton built on the following set of words X

$$X = \{u_i \cdot w \mid 1 \leq i \leq r \text{ and } w \in \{\epsilon\} \cup \mathcal{E}_{i,j} \text{ for some } j\}.$$

The automaton \mathcal{T} is built on X with $Q = \text{Pref}(X)$ (set of states), $i = \epsilon$ (initial state). The transition function is defined (as in Aho-Corasick construction) as

$$\delta(p, x) = \text{the longest suffix of } px \in \text{Pref}(X).$$

In order to count the number of clumps (for instance) the set of final states T needs more attention: it is defined as

$$T = X \setminus X\mathcal{A}^+.$$

This automaton accepts the language of words ending by the first occurrence of a word in a clump.

We can easily derive from this automaton the generating function $f(z, x_1, \dots, x_r, t, u)$ where x_i marks an occurrence of u_i , t marks the number of clumps, and u the total length covered by the clump. Indeed, one has to mark some transitions in the adjacency matrix A according to some simple rules.

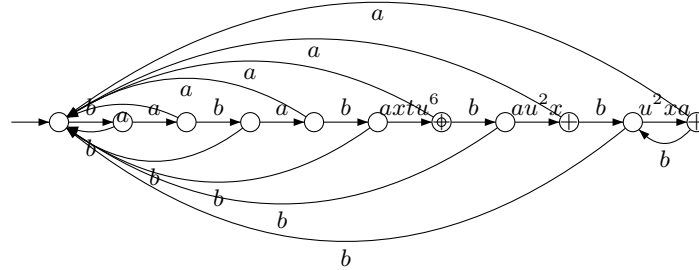
- To count occurrences of the u_i 's, we have to mark with the formal variable x_i transitions going to states $\mathcal{A}^*u_i \cap \text{Pref}(X)$ (for $1 \leq i \leq r$).
- For the number of clumps, one can mark transitions going to states in $\mathcal{U} \setminus \mathcal{U}\mathcal{A}^+ = X \setminus X\mathcal{A}^+$, that is states corresponding to first occurrences inside a clump.
- Finally, for the total length covered by clumps. We have to put a formal weight on transitions going to a state $p \in \mathcal{A}^*\mathcal{U} \cap \text{Pref}(X)$ taking into account the number of symbols between the last occurrence of a word of X and the new one at the end of p . Let us define for a state p (corresponding to a word with a occurrence of some word of X at the end) the function $\ell(p)$ the maximal proper prefix q of p in $\mathcal{A}^*\mathcal{U}$ if it exists or ϵ if there is no such prefix. Then we must mark all transitions going to p with $u^{|p| - |\ell(p)|}$ (if $p \in \mathcal{A}^*\mathcal{U} \cap \text{Pref}(X)$).

Of course the construction does not gives a minimal automaton. However the automaton is complete and deterministic so that the translation to generating function is straightforward.

Example

1. For one word $\mathcal{U} = \{u = bababa\}$, and $\mathcal{E}_u = \{ba, baba\}$. The set X is

$$X = \{bababa, babababa, bababababa\}.$$



N.B.: The sign '+' on the automaton indicates that the corresponding prefix ends with some occurrence of \mathcal{U} . The double oval states indicates the states where we know we have entered a new clump.

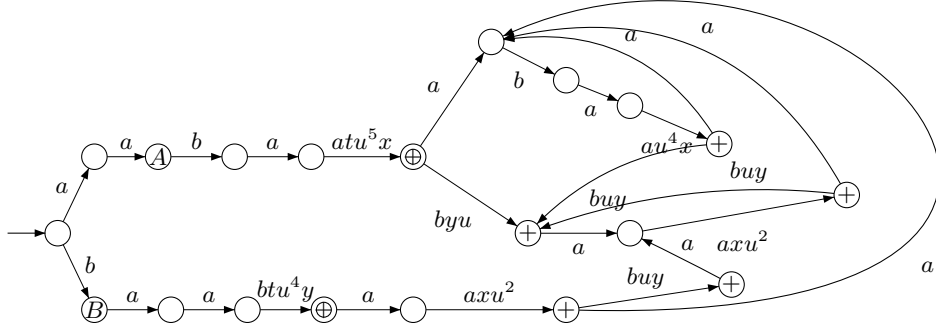
2. For the set $\mathcal{U} = \{u_1 = aabaa, u_2 = baab\}$ and the matrix of right extension sets is

$$\mathcal{E} = \begin{pmatrix} baa + abaa & b \\ aa & aab \end{pmatrix}.$$

The set X is

$$X = \{aabaa, aabaab, aabaabaa, aabaaabaa, baab, baabaa, baabaab\}.$$

We have the following automaton (with x and y marking occurrences respectively of u_1 and u_2 . The automaton is complete and deterministic. However, for clarity's sake, all transitions labelled by a and b ending respectively on state A and B are omitted. As before, the sign '+' indicates that the corresponding prefix (or, equivalently, state) ends with some occurrence of \mathcal{U} . The double oval attribute indicates the state where we know we have entered a new clump.



5 Limit laws

5.1 Normal law

A normal limit law for the number of clumps U when $U = O(n)$ in texts of size n follows from the automaton construction of Section 4. A Perron-Frobenius property asserts the existence of a unique dominant eigenvalue of the positive system; apply next a suitable Cauchy integral and large power Theorem of Hwang [12, 13]; see [16] for details.

5.2 Poisson law for rare words

In a Bernoulli model, if \underline{p} and \bar{p} are the minimal and maximal probability of letters of the alphabet, words of size $l < \frac{\log n}{\log(1/\bar{q})}$ have $O(n)$ number of occurrences in texts of size n with probability one. We consider rare words with size over this threshold and number of occurrences $O(1)$. We prove in this case a Poisson-like limit law. Taking a Taylor expansion of $O^{(\gamma)}(z, u)$ in Equation (12) at $u = 0$, and considering the k th Taylor coefficient, with $k = O(1)$ provide a rational generating function with respect to the variable z of the form

$$H_k(z) = [u^k]O^{(\gamma)}(z, u) = \frac{\mathcal{R}(z)\mathcal{U}(z)(\mathcal{M}(z) - \mathcal{K}(z))^{k-1}}{(1 - \mathcal{K}(z))^k} = \frac{\pi_w z^{|w|} (z - 1 + (1 - \mathcal{K}(z))D(z))^{k-1}}{(1 - \mathcal{K}(z))^k (D(z))^{k+1}}. \quad (14)$$

We follow Fayolle [6] to prove that the dominant root of the denominator of this last equation is the smallest and positive root of $D(z) = \pi_w z^{|w|} + (1 - z)C(z)$; (see Equation (5)). Let d be the smallest period of w . If $d \leq l/2$ classical results about periods on words provide $C(z) = 1 + \pi_u z^{|u|} + \dots + (\pi_u z^{|u|})^r + S(z)$ for a given word u with $|u| < l/2$, and $r \geq 2$; moreover $S(z)$ is a polynomial of minimal degree at least $l/2$. Moreover, we have $\mathcal{K}(z) = \pi_u z^{|u|} + R(z)$ where $S(z) - R(z)$ is a polynomial with positive coefficients. This entails that $|S(z)|$ and $|R(z)|$ are $o(1)$ for $|z| < 1/\bar{p}$. Up to negligible terms, we get

$$|C(z)| = \left| \frac{1}{1 - \pi_u z^{|u|}} \right| \geq \frac{1}{1 + \pi_u |z|^{|u|}} \geq \frac{1}{1 + p|z|} \quad \text{for } |z| < \frac{1}{p}.$$

We also have $|1 - \mathcal{K}(z)| > 0$ and $\pi_w z^{|w|} = o(1)$ for $|z| < 1/p$. The Rouché theorem in the disk $|z| < 1/p$ the generating function $H_k(z)$ has a single pole which is a smallest modulus root ρ of the equation $D(z) = 0$. Perron-Frobenius considerations on the automaton counting the number of occurrences of w imply that this pole is real positive. A similar proof follows when $d > l/2$.

Writing $D(z) = Q(z)(1 - z/\rho)$ and $P(z) = z - 1 + (1 - \mathcal{K}(z))D(z)$ we get as a first approximation

$$\mathbf{P}(O_n^\gamma = k) \approx \frac{\pi_w \rho^{|w|}}{Q(\rho)} \times \frac{1}{k!} \left(\frac{\rho P(\rho) \times n}{(1 - \mathcal{K}(\rho))Q(\rho)} \right)^k \times \rho^{-n}.$$

A similar behaviour has been observed for occurrences of one word by Régnier and Szpankowski [18].

5.3 Length of the clumps in infinite texts

Generating function of the size of the clumps in infinite texts is a sum of geometric random variables.

6 Conclusion

An interesting application of this article would be a combinatorial analysis of *tandem repeats* or multiple repeats that occur in genomes; large variations of such repeats are characteristics of some genetic diseases.

Would it be possible to extend our approach to clumps of regular expressions? We consider *clumps of a regular expression* (i.e. contiguous sets of positions such that each position is covered by at least one word of the associated regular language and such that leading and terminating positions of each occurrence is covered by at least two occurrences). In this case the star-height theorem (CITE) implies that we cannot in general find a finite set of words w_i and a finite set of prefix codes \mathcal{K}_i with $1 \leq i \leq \ell$ such that the language $\bigcup_{1 \leq i \leq \ell} w_i(\mathcal{K}_i)^*$ describes the clumps.

References

- [1] BARBOUR, A., HOLST, L., AND JANSON, S. *Poisson Approximation*. Oxford University Press, 1992.
- [2] BASSINO, F., CLÉMENT, J., FAYOLLE, J., AND NICODÈME, P. Counting occurrences for a finite set of words: an inclusion-exclusion approach. In *Proceedings of the 2007 Conference on Analysis of Algorithms* (2007), P. Jacquet, Ed., DMTCS, proc. AH, pp. 29–44. Proceedings of a colloquium organized by Juan-les-Pins, France, June 2007.
- [3] BERSTEL, J. Growth of repetition-free words - a review. *Theoretical Computer Science*, 340 (2005), 280–290.
- [4] BERSTEL, J., AND PERRIN, D. *Theory of Codes*. Pure and Applied Mathematics. Academic Press, 1985.
- [5] CHOMSKY, N., AND SCHÜTZENBERGER, M. The algebraic theory of context-free languages. *Computer Programming and Formal Languages*, (1963), 118–161. P. Braffort and D. Hirschberg, eds, North Holland.
- [6] FAYOLLE, J. An average-case analysis of basic parameters of the suffix tree. In *Mathematics and Computer Science* (2004), M. Drmota, P. Flajolet, D. Gardy, and B. Gittenberger, Eds., Birkhäuser, pp. 217–227. Proceedings of a colloquium organized by TU Wien, Vienna, Austria, September 2004.
- [7] GANTMACHER, F. *The theory of matrices. Vols. 1,2*. Encyclopedia of Mathematics. New York: Chelsea Publishing Co. Translated by K. A. Hirsch, 1959.

- [8] GOULDEN, I., AND JACKSON, D. An inversion theorem for clusters decompositions of sequences with distinguished subsequences. *J. London Math. Soc.* 2, 20 (1979), 567–576.
- [9] GOULDEN, I., AND JACKSON, D. *Combinatorial Enumeration*. John Wiley, New York, 1983.
- [10] GUIBAS, L., AND ODLYZKO, A. Periods in strings. *J. Combin. Theory A*, 30 (1981), 19–42.
- [11] GUIBAS, L., AND ODLYZKO, A. Strings overlaps, pattern matching, and non-transitive games. *J. Combin. Theory A*, 30 (1981), 108–203.
- [12] HWANG, H.-K. *Théorèmes limites pour les structures combinatoires et les fonctions arithmétiques*. PhD thesis, École polytechnique, Palaiseau, France, Dec. 1994.
- [13] HWANG, H.-K. Large deviations for combinatorial distributions I: Central limit theorems. *Ann in Appl. Probab.* 6 (1996), 297–319.
- [14] JACQUET, P., AND SZPANKOWSKI, W. Autocorrelation on words and its applications. Analysis of Suffix Trees by String Ruler Approach. *J. Combin. Theory A*, 66 (1994), 237–269.
- [15] LOTHAIRE, M. *Applied Combinatorics on Words*. Encyclopedia of Mathematics. Cambridge University Press, 2005.
- [16] NICODÈME, P., SALVY, B., AND FLAJOLET, P. Motif statistics. *Theoretical Computer Science* 287, 2 (2002), 593–618.
- [17] RÉGNIER, M. A unified approach to word occurrences probabilities. *Discrete Applied Mathematics* 104, 1 (2000), 259–280. Special issue on Computational Biology.
- [18] RÉGNIER, M., AND SZPANKOWSKI, W. On pattern frequency occurrences in a markovian sequence? *Algorithmica* 22, 4 (1998), 631–649. This paper was presented in part at the 1997 International Symposium on Information Theory, Ulm, Germany.
- [19] REINERT, G., AND SCHBATH. Compound poisson and poisson process approximations for occurrences of multiple words in markov chains. *J. Comp. Biol.* 5 (1998), 223–253.
- [20] REINERT, G., SCHBATH, S., AND WATERMAN, M. Probabilistic and statistical properties of words: an overview. *J. Comp. Biol.* 7 (2000), 1–46.
- [21] RIVALS, E., AND RAHMANN, S. Combinatorics of periods in strings. *Journal of Combinatorial Theory - Series A* 104, 1 (2003), 95–113.
- [22] SCHBATH, S. Compound poisson approximation of word counts in DNA sequences. *ESAIM Probab. Statist.* 1 (1995), 1–16.
- [23] SEDGEWICK, R., AND FLAJOLET, P. *An Introduction to the Analysis of Algorithms*. Addison-Wesley Publishing Company, 1996.
- [24] STEFANOV, V., ROBIN, S., AND SCHBATH, S. Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Applied Mathematics* 155 (2007), 868–880.
- [25] SZPANKOWSKI, W. *Average Case Analysis of Algorithms on Sequences*. Series in Discrete Mathematics and Optimization. John Wiley & Sons, 2001.